DOCUMENT RESUME

ED 295 953 TM 011 707

AUTHOR Johnson, Eugene
TITLE Considerations and Techniques for the Analysis of NAEP Data.
PUB DATE Apr 88
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Data Analysis; Elementary Secondary Education; Mathematics Tests; *Sampling; Scores; *Standardized Tests; *Test Validity
IDENTIFIERS Multiple Matrix Sampling; *National Assessment of Educational Progress; Plausibility (Tests); Scaling Models; Stratified Sampling

ABSTRACT

The special characteristics of the data from the National Assessment of Educational Progress (NAEP) that affect the validity of conventional techniques of statistical analysis are considered. In contrast to the assumptions underlying standard methods of statistical analysis, the NAEP samples are obtained via a stratified multi-stage probability sampling design in which clusters of students are selected and in which certain sub-populations are sampled at higher rates than others. The resulting samples have different statistical characteristics than those of a simple random sample. Analytic techniques that take the structure of the data into account are discussed, and their properties are explored. A second feature of the NAEP data arises from the fact that multiple matrix sampling techniques are used for the assignment of sub-sets of the pool of exercises to sub-samples of students. The NAEP accounts for the effect of the missing data on its scaling models through the use of multiple imputation procedures; creating vectors of "plausible values" for the scale scores. The use of plausible values and their effect on analysis are illustrated. The analyses are illustrated with results from the NAEP assessments of mathematics. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295953

CONSIDERATIONS AND TECHNIQUES FOR THE ANALYSIS OF NAEP DATA

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Eugene Johnson

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
EUGENE JOHNSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

The special characteristics of the National Assessment data which affect the validity of conventional techniques of statistical analysis are considered. In contrast to the assumptions underlying standard methods of statistical analysis, the NAEP samples are obtained via a stratified multi-stage probability sampling design in which clusters of students are selected and in which certain subpopulations are sampled at a higher rate. The resulting samples have different statistical characteristics than those of a simple random sample. Analytic techniques which take the structure of the data into account are discussed and their properties explored. A second feature of the National Assessment data arises from the fact that multiple matrix sampling techniques are used for the assignment of subsets of the pool of exercises to subsamples of students. National Assessment accounts for the effect of the missing data on its scaling models through the use of multiple imputation procedures; creating vectors of "plausible values" for the scale-scores. The use of plausible values and their effect on analysis are illustrated. The analyses are illustrated with results from the NAEP assessments of mathematics.

Presented at the Annual Meeting of the American Educational Research Association in New Orleans, April 5-9, 1988.

TM011707

CONSIDERATIONS AND TECHNIQUES FOR THE ANALYSIS OF NAEP DATA

Eugene Johnson [1]

INTRODUCTION

The primary goals of the National Assessment of Educational Progress are to detect and report the current status of, as well as changes in, the educational attainments of young Americans. To accomplish these goals, NAEP biennially selects large, nationally representative, samples of students and from these samples gathers a vast quantity of information about the knowledge, skills, understandings, backgrounds, and attitudes of these students in major learning areas usually taught in school.

The purpose of this paper is to consider the special characteristics of the National Assessment data which affect the validity of conventional techniques of statistical analysis.

Many standard methods of statistical analysis are based on the assumption that the data have been collected by a simple random sample of the population of interest. In contrast, the NAEP samples are obtained via a stratified multi-stage probability sampling design in which clusters of students are selected and in which certain subpopulations are sampled at a higher rate. The resulting samples have different statistical characteristics than those of a simple random sample. Analytic techniques which take the structure of the data into account will be discussed and their properties explored. In particular, the consequences of ignoring the structure of the data and proceeding with classical techniques will be examined. These analyses will be illustrated with results from the NAEP assessments of mathematics.

A second feature of the National Assessment data arises from the fact that, due to limitations on the amount of student time available and due to considerations of statistical efficiency, multiple matrix sampling techniques are used for the assignment of subsets of the pool of exercises to subsamples of students. The result is that a substantial amount of data is missing by design. National Assessment uses scaling models to summarize the performance of students in a learning area or subarea and to summarize the responses of students to sets of background questions. In both cases, these summaries are possible even though different students have responded to different sets of questions. National Assessment accounts for the substantial amount of missing data by using multiple imputation procedures; creating vectors of "plausible values" for the scale-scores. The use of plausible values and their effect on analysis will be illustrated with results of an analysis based on the plausible values of the Mathematics Composite, an overall measure of Mathematics achievement.

---

[1] The statistical programming for this paper was performed by David Freund, Laura McCamley and Jennifer Nelson.

1. THE EFFECT OF THE SAMPLE DESIGN ON ANALYSIS

The goal of the 1986 National Assessment sample design was to obtain and assess a sample of students in order to derive various estimates of population and subpopulation characteristics (e.g. average Mathematics achievement for the population of in-school students in the eleventh grade; average achievement for 11th grade students in the Northeast). To be useful, these estimates were required to be consistent, in the sense that they appropriately represent the population values, and to have reasonably high precision, in the sense that the estimates have tolerably low sampling variability. Balancing this requirement for precision was the necessity that the sample be economically and operationally feasible to obtain.

As was the case in previous assessments, the 1986 National Assessment used a complex sample design to balance the conflicting goals of high precision versus economic and operational feasibility. The target population for the 1986 Assessment consisted of 9-year-olds, 13-year-olds, and 17-year-olds enrolled in public and private elementary and secondary schools, along with other students in the modal grade for each of these three ages. The sample was obtained from this target population using a multistage probability design with four stages of selection. The first stage of selection, the primary sampling units (PSUs), consisted of counties or groups of counties. The second stage of selection consisted of elementary and secondary schools within the selected PSUs. The assignment of assessment sessions to sampled schools comprised the third stage of sampling, and the fourth stage involved the selection of students within schools and their assignment to the sessions.

The probabilities of selection of the first- and second-stage sampling units were proportional to measures of their size so that larger PSUs and schools were more likely to be selected into the sample than were smaller PSUs and schools. The probabilities for subsequent stages of selection were determined so that the overall probabilities of selection of students were, in general, approximately uniform, with exceptions for certain subpopulations that were oversampled by design. Schools with high concentrations of Black, Hispanic, or Asian students were deliberately sampled at twice the normal rate to obtain larger samples of respondents in order to increase the precision in the estimation of the characteristics of these subpopulations.

The resulting sample has characteristics which are very different from those of a simple random sample in which every student in the target population has an equal chance of selection and in which the observations from different sampled students can be considered statistically independent of one another.

An obvious difference between the NAEP sample and a simple random sample arises because the NAEP sample design uses differential sampling rates so that certain demographic subgroups of students are overrepresented in the sample. That is, the achieved sample contains proportionately more members of these oversampled subgroups than there are in the population. Appropriate estimation of population characteristics must take this disproportionate representation into account. This is accomplished by assigning a weight to

each respondent, where the weights properly account for the sample design and reflect the appropriate proportional representation of the various types of individuals in the population. The next section will consider the effect of ignoring the sampling weights in the analysis of the NAEP data.

Another consequence of the NAEP sample design is its effect on the estimation of sampling variability. Because of the effects of cluster selection (students within schools, schools within Primary Sampling Units) and because of the effects of certain adjustments to the sampling weights (non-response and post-stratification), observations made on different students cannot be assumed to be independent of one another. As a result, ordinary formulae for the estimation of the variance of sample statistics, based on assumptions of independence and homoscedasticity, will tend to underestimate the true sampling variability. Furthermore, the degrees of freedom attributable to an estimate of variability are always less for the NAEP data than they would be for a simple random sample of the same size. These two effects taken together mean that conventional (i.e. simple random sampling based) inferential techniques will tend to underestimate the error variance and overestimate its degrees of freedom, resulting in statistical tests that are too liberal in that "too many" statistically significant results will be obtained. These points will be discussed further in later sections, in which the effects of ignoring the sample design on variance estimation will be explored.

1.1 Effect of Weighting on Estimation

As observed above, the deliberate oversampling of certain subpopulations in order to enhance the precision estimates for these oversampled subgroups produces an achieved sample containing proportionately more members of these subgroups than there are in the population. In addition to the deliberate overrepresentation of certain subgroups from oversampling, nonresponse on the part of schools and students results in a final sample with too many of some types of students, and too few of others, relative to what would be found in the target population. To bring the proportional representation of the various subgroups in the sample into line with the population proportions, NAEP assigns a weight to each assessed student.

The weight assigned to a given student is the product of a number of components, each of which is designed to adjust for a particular characteristic of the achieved sample. The major component of the student's weight is the student's base weight, which is computed as the reciprocal of the overall probability that the student was selected for assessment and which therefore adjusts for the effects of disproportionate sampling. These base weights were then adjusted for nonresponse at the school and student level. Finally, the weights were adjusted, by post-stratification, to make sample estimates of certain subpopulation totals (defined in terms of race, ethnicity, census region, and the size and degree of urbanization of a county) conform to external, more accurate, estimates based on Census and Current Population Survey data.

The final adjusted weight assigned to a student properly reflects the sample design, adjusts for nonresponse, and reflects, as much as is possible, the proper representation of the various types of individuals in the population. This weight was used to derive all estimates of population and subpopulation characteristics which have been presented in the various NAEP reports, including simple estimates such as the proportion of students of a specified type who would respond in a certain way to an exercise and more complex estimates such as mean proficiency levels.

Although it is straightforward to use the weights within standard statistical packages to get properly weighted estimates of statistics such as means and regression coefficients, it is interesting to see how much difference the use of weights makes in the final results. We shall examine the impact of the sampling weights on two measures of achievement from the 1986 Assessment of Mathematics, item level percent correct statistics and proficiency scale-scores. We shall restrict our attention to statistics measuring the Mathematics achievement of students in the eleventh grade.

A basic measure of achievement in Mathematics occurs at the level of an individual item and consists of the percent of students in the population of interest predicted to answer the item correctly. This item-level percent correct is properly estimated by P_W, 100 times the ratio of the sum of the weights of all sampled students in the population of interest who responded correctly to the item to the sum of the weights of all sampled students in the population who responded to the item (but not necessarily correctly). The unweighted analogue of P_W is P_U, 100 times the number of students answering correctly divided by the total number responding to the item.

Table 1 shows how different P_W can be from P_U by showing elements of the distribution of the difference (P_U-P_W) across the set of 372 cognitive Mathematics items presented to students in the eleventh grade. Included in the table are the mean difference, the median difference, and the upper quartile (HiQ) and the lower quartile (LoQ) of the distribution of differences between the unweighted and weighted percents correct. These statistics are shown for the full population of eleventh grade students ("TOTAL") as well as for the following selected demographic subgroups of the students: Males, Females, White Students, Black Students, Hispanic Students, Students in Disadvantaged Urban Environments ("Low Metropolitan"), and Students in Advantaged Urban Environments ("High Metropolitan").

For practical purposes, the weighted percent correct statistic P_W provides an unbiased estimate of the percent of students in the entire population of interest who would respond correctly to the item. (The statistic P_W technically has a bias of the order of $1/n$, where n is the sample size, since P_W is the ratio of two random variables -- because the sample size is large for most subgroups of interest, this technical bias can be ignored).

The difference (P_U-P_W) is thus interpretable as the bias incurred by using the unweighted percent correct statistic to estimate the population percent correct on the item. From Table 1 we see that the unweighted percent correct tends to <u>underestimate</u> the population value for the total population

of eleventh grade students, as well as for the subpopulations of Males, Females, Blacks, and students in the High Metropolitan environment. Conversely, P_U tends to <u>overestimate</u> the percent correct for White and Hispanic students, and, to a much lesser extent, for students in the Low Metropolitan environment. P_U will have a bias to the extent that the magnitude of the weights are related to performance on the item. Thus, a negative bias occurs when the sample of the subpopulation of interest contains too many lower performing students which are adjusted for in the weighted sample by having relatively smaller weights. A positive bias will occur when there are either too many high performing students or too few low performing students (or, perhaps, both) in the sample.

Table 1

Distribution by Demographic Subgroup of the Difference Between the Unweighted and Weighted Percent Correct Statistics for the 372 Mathematics Items Presented at Grade 11

Group	Mean	HiQ	Median	LoQ
Male	-.51	.13	-.48	-1.16
Female	-.88	-.18	-.78	-1.50
White	.21	.75	.22	-.28
Black	-.61	.45	-.61	-1.67
Hispanic	.24	1.82	.34	-1.10
Low Metro	.03	1.80	.07	-1.48
High Metro	-1.29	-.11	-1.28	-2.63
TOTAL	-.74	-.16	-.60	-1.32

Before leaving this example, it is germane to note that the biases reported in Table 1 tend to be no larger than the standard errors of the percent correct statistics. The standard errors of the mean of the 372 weighted percent correct statistics are between .6 and .8 for Whites, Males, Females, and the total population; and are between 1.4 and 2.3 for the remaining subpopulations shown in the table. The standard errors for an individual item's percent correct statistic are 1.3 to 1.9 for Whites, Males, Females, and total; and 3 to 5 for the remaining subpopulations.

Consequently, in one sense, the effects of the biases due to not weighting in the estimation of item level percent correct statistics are reduced when considered in the context of the sampling errors of the statistics. In another sense, the use of unweighted statistics introduces a bias that can easily be eliminated by the use of the sampling weights.

Since most of the analyses conducted by NAEP are based on the results of scaling models which summarize performance of students across a learning area, such as Mathematics, we next consider the effects of weighting on analyses based on these scale-scores. For this purpose, we shall consider the Mathematics Composite scale-score, an overall Mathematical proficiency measure, which will be described in Section 2. As we shall note there, the relatively small number of items answered by any individual precludes the use of the traditional point estimates of proficiency available from traditional IRT applications. Rather, NAEP provides each individual with a set of "plausible values", each of which is a random draw from the distribution of the potential scale-scores for that individual. Since our current interest is on the effect of the sampling design on estimation and inference, we will restrict our attention to a single measure of an individual's Mathematical proficiency, the first plausible value of the individual's Mathematics Composite scale-score.

To assess the potential effect of weighting on more complex analyses of NAEP data, we compare the results of two regression analyses, one weighted and one unweighted, each of which addresses the question of whether the linear relationship between Mathematical proficiency and the highest level of mathematics course taken is the same for White students and Black students. The equation for both regressions is as follows:

$$Mpv1 = \alpha + \beta*BW + \gamma*NMATH + \beta\gamma*BW*NMATH$$

where

Mpv1 is the first plausible value of the student's Composite Mathematics scale-score,

BW is the student's race coded 1=White, -1=Black

NMATH is the highest Mathematics course taken, coded
1=Pre-Algebra
2=Algebra
3=Geometry
4=Algebra 2
5=Calculus.

BW*NMATH addresses the difference in slopes of Mpv1 on NMATH between White students and Black students.

Table 2 shows the resultant estimates of the regression coefficients for the set of 10105 students in the eleventh grade who were either white or black and who were provided a score on the Mathematics composite. The first panel of the table shows the unweighted estimates, with standard errors in parentheses, and the second column provides the weighted estimates, also with standard errors. The fit of these regressions is reasonably good; the multiple correlation being .75 in both cases.

It should be noted that the standard errors supplied in this table are computed under the assumption of simple random sampling and so, as we shall see in the next section, underestimate the true variability of the regression coefficients. It should also be noted that, for the weighted analysis, the

weights have been scaled to add to the total sample size of 10105, rather than to their original total of 1,060,564, which is an estimate of the number of White and Black eleventh grade students in the population. By scaling the weights to add to the sample size, we avoid a clear overestimation of the real amount of information available in the construction of the standard errors. We shall see in the next section that this scaling does not go far enough in correctly representing the real amount of information available.

Table 2
Coefficients of the Unweighted and Weighted Regressions of Mpv1 on Race and Highest Math Course Taken
(simple random sampling standard errors in parentheses)

Coefficient		Unweighted Estimate		Weighted Estimate		Difference
Intercept	(α)	253.92	(0.59)	252.80	(0.64)	1.12
BW	(β)	6.95	(0.59)	7.75	(0.64)	-.80
NMATH	(γ)	14.01	(0.20)	14.54	(0.21)	-.53
BW*NMATH	($\beta\gamma$)	1.73	(0.20)	1.28	(0.21)	.45

Analogous to the percent correct statistics, the difference between the unweighted regression coefficients and the weighted estimates provides a measure of the bias incurred as a result of neglecting the sampling weights. We see from Table 2 that this bias is positive for the intercept and the BW by NMATH interaction and negative for the BW effect and the slope on NMATH. Nearly all of the bias in the unweighted coefficients comes from the improper treatment of the regression of Mpv1 on NMATH for the Black students.

Table 3 shows the coefficients of the unweighted and weighted regressions of Mpv1 on NMATH separately for the White and the Black students. It can be seen that the difference between the two regressions for the White students is relatively slight, with the difference between the unweighted and weighted coefficients being about half a standard error. On the other hand, the unweighted regression line for the Black students is noticeably flatter than its weighted counterpart with the unweighted slope being about 2.5 nominal (Simple Random Sampling) standard errors smaller than the weighted slope. The effect of weighting is greater for the Black students than the White students because relatively more of the Black students were subject to differential weighting (due primarily to oversampling) than were the White students. Coupled with this is an apparent relationship between the size of a student's weight and the level of proficiency, conditional on the highest Mathematics course taken (so that, for a given level of NMATH, students with smaller weights -- who are overrepresented in the sample -- tend to have lower values of Mpv1). The result is that the unweighted regression understates the relationship between NMATH and Mpv1 for Black students.

Table 3
Coefficients by Race of the Unweighted and Weighted
Regressions of Mpv1 on NMATH
(simple random sampling standard errors in parentheses)

Race	Coefficient	Unweighted Estimate		Weighted Estimate		Difference
White	Intercept	260.87	(0.50)	260.55	(0.54)	.32
	Slope	15.74	(0.14)	15.82	(0.15)	-.08
Black	Intercept	246.97	(1.07)	245.05	(1.16)	1.92
	Slope	12.28	(0.37)	13.26	(0.39)	-.98

The bottom line from our limited comparisons of the effects of weighting on the estimation of sample statistics is that the use of weights does affect the results but that often the consequences of not weighting is moderate since, particularly in the case of percent correct statistics, the bias incurred due to neglecting the weights is generally no larger than the apparent variability of the statistic being estimated. The regression example showed that, in more complex situations, weighting can become important. The consequences of not weighting are likely to increase as the complexity of the analysis increases. Because of this, and because modern statistical packages readily allow the use of weights in analysis, there is no compelling reason to accept the biases due to ignoring weights.

Much more difficult than the elimination of bias due to not weighting, is the task of properly accounting for the variability of the resultant statistics, a topic to which we now turn.

1.2 The effect of Sample Design on the Estimation of Sampling Variability

A major source of uncertainty in the estimation of the value in the population of a variable of interest exists because information about the variable is obtained on only a sample from the population. To reflect this fact, it is important to attach to any statistic (e.g., a mean) an estimate of the sampling variability to be expected for that statistic. Estimates of sampling variability provide information about how much the value of a given statistic would be likely to change if the statistic had been based on another, equivalent, sample of individuals drawn in exactly the same manner as the achieved sample. Consequently, the estimation of the sampling variability of any statistic must take into account the sample design.

As we have noted before, the NAEP sample is obtained via a stratified multistage probability sampling design which includes provisions for sampling certain subpopulations at higher rates. Additional characteristics of the

sample include adjustments for both nonresponse and post-stratification. The resulting sample has different statistical characteristics from those of a simple random sample. In particular, because of the effects of cluster selection (students within schools, schools within PSUs) and because of effects of nonresponse and post-stratification adjustments, observations made on different students cannot be assumed to be independent of each other (and are, in fact, generally positively correlated). Furthermore, to account for the differential probabilities of selection (and the various adjustments), each student has an associated sampling weight, which should be used in the computation of any statistic and which is itself subject to sampling variability. Treatment of the data as a simple random sample, with disregard for the special characteristics of the NAEP sample design, will produce underestimates of the true sampling variability.

It is unfortunately the case that the proper estimation of the sampling variability of a statistic based on the NAEP data is complicated and requires techniques beyond those commonly available in standard statistical packages. Fortunately, a procedure exists, the jackknife, which provides good quality estimates of the sampling variability of any statistic, at the expense of increased computation, and which can be used in concert with standard statistical packages to obtain a proper estimate of sampling variability.

The jackknife procedure used by NAEP has a number of properties which make it particularly suited for the analysis of NAEP data:

1) It properly estimates the sampling error arising from the complex sample selection procedure for simple estimates such as totals, and does so approximately for more complex estimates.

2) It reflects the component of sampling error introduced by the use of weighting factors which are dependent upon the sample data actually obtained.

3) It can be adapted readily to the estimation of sampling errors for parameters estimated using statistical modeling procedures, as well as for tabulation estimates such as totals and means.

4) Once appropriate weights are derived and attached to each record, jackknifing is straightforward to use for estimating sampling errors. A single set of replicate weights is required for all tabulations and model parameter estimates which may be needed.

The method of applying the jackknife procedure involves first defining pairs of groups of first-stage sampling units (FSSUs, which are, roughly, PSUs). For the 1986 NAEP main assessment, 76 groups of such first-stage units were defined. These FSSUs were then paired so that the populations represented by each member of the pair are similar, resulting in 38 pairs of first-stage units. Estimation of the sampling variability of a statistic t is then based on the estimation of the variance between first-stage units (within strata) of the sample estimates for these units. Appropriate estimation by this approach will reflect approximately the combined effect of the between- and within-FSSU contributions to variance because of the

remarkable and convenient fact that, under conditions satisfied by the NAEP sample, variance estimates based on the differences between estimates at the level of the first-stage units also appropriately reflect the variability within those units, no matter how the subsampling within FSSUs was done. (For a discussion see Wolter, 1985, Section 2.4.5.; Hansen, Hurwitz, & Madow, 1953, p. 258.)

The jackknife method estimates the sampling variability of any statistic as the sum of components of variability which may be attributed to each of the FSSU pairs. The variance attributed to a particular FSSU pair is measured by estimating how much the value of the statistic would change if the information embodied in the FSSU pair were to be changed. This is done by the computation of a quantity t_i called a pseudoreplicate, which is associated with the i^{th} FSSU pair, and which is an estimate of the statistic of interest t based on an altered sample. Specifically, the i^{th} pseudo-replicate of the statistic t is created by eliminating the data from the first FSSU of the pair, replacing the lost information with that from the second FSSU of the pair (so that the second FSSU is included twice), and then re-estimating the statistic based on this altered set of data.

The component of the sampling variability attributable to a pair of first-stage units is estimated as the squared difference between the value of the statistic for the complete sample and the pseudoreplicate associated with the pair. For computational purposes, the pseudoreplicate associated with a given pair is the original statistic computed with a different set of weights, referred to as the student replicate weights, which we shall denote as SRWT01 through SRWT38, where SRWTi is for the i^{th} pair. This set of weights allows measurement of the total effect of replacing one member of the pair with a copy of the other, including adjustments for nonresponse and post-stratification. The pseudoreplicate associated with the i^{th} pair for a given statistic is obtained by recalculating the statistic using the weights SRWTi instead of the sampling weights.

As a specific example of the use of the student replicate weights, let $t(\underline{y},\underline{w})$ be any statistic which is a function of the sample responses $\underline{y}$ and the weights $\underline{w}$ and which estimates population value T. For example, t could be a weighted mean, a weighted percent-correct point or a weighted regression coefficient. The $t(\underline{y},\underline{w})$, computed with the sampling weights is the appropriate sample estimate of T. To compute $\hat{Var}(t)$, the sampling variance for this statistic, proceed in the following manner:

1) For each of the 38 pairs of first-stage units, compute the associated pseudoreplicate for the statistic. For the i^{th} pair, this is

$$t_i = t(\underline{y},\underline{SRWTi})$$

which is the statistic t recalculated by using SRWTi instead of the sampling weights.

2) The sample variance of t is

$$\hat{Var}(t) = \sum_{i=1}^{38} (t_i - t)^2 \quad .$$

In practical terms, the major expenditure of resources in the computation of a jackknife variance estimate occurs in the construction of the pseudoreplicates. In the 1986 assessment, this implies that the statistic of interest has to be recomputed 39 times, once for the overall estimate t, and once for each of the 38 pseudoreplicates t_i. Because this is a considerable increase in the amount of computation required, relative to a conventional variance estimate, it is of interest to see how much the jackknife variance estimates differ from their less computationally intensive, simple random sampling based, analogues. For this purpose, we will compare the jackknife variance estimates with the conventional estimates for the item percent correct statistics and for the regression coefficients which were considered in the last subsection.

The comparison of the two methods of variance estimation will be in terms of a statistic called the <u>design effect</u>, which was developed by Kish (1965) and extended by Kish and Frankel (1974). The design effect for a statistic is the ratio of the actual variance of the statistic (taking the sample design into account) over the conventional variance estimate based on a simple random sample with the same number of elements. To avoid sources of bias due to improper representation, this conventional estimate uses the sampling weights. The design effect is the inflation factor to be applied to the conventional variance estimate in order to adjust error estimates based on simple random sampling assumptions to account approximately for the effect of the sample design. The value of the design effect depends on the type of statistic computed and the variables considered in a particular analysis as well as the combined clustering, stratification, and weighting effects occurring among sampled elements. Generally, the design effects for statistics from complex samples such as NAEP are greater than one, causing variances based on simple random sampling assumptions to tend to be underestimates.

As an example of the distribution of design effects to be expected from NAEP data, we consider the design effects for the weighted percent correct statistics for the cognitive Mathematics items. The design effect for the weighted percent, P, of correct responses to a given item is of the form

$$deff(P) = Var_{JK}(P)/(P(100 - P)/N) \quad .$$

In the above, N is the total number of individuals in the subgroup responding to the exercise, $Var_{JK}(P)$ is the jackknife variance of P, and

P(100 - P)/N is the conventional variance estimate of P. (Although the estimate P(100 - P)/N has the same form as the simple random sampling estimator of the variance of a percent correct, the sample weights have been taken into account via the weighted estimation of P.)

Table 4 addresses the distributions of the design effects for the 372 cognitive Mathematics items presented to Grade 11 students. These distributions are shown for the population as a whole ("total") as well as for a variety of demographic subgroups: sex; race/ethnicity (White, Black, Hispanic, other); age (less than modal age, modal age, greater than modal age); region (Northeast, Southeast, Central, West); Size and Type of Community (Rural, Low Metropolitan, High Metropolitan, Big City, Urban Fringe, Medium, City, Small Place); and parental education (At Most High School, Graduated High School, Post-High School, Graduated College, Unknown). For each of these groupings of Grade 3 students, Table 4 provides the lower quartile (LoQ), median, upper quartile (HiQ) and maximum design effect as well as the mean design effect.

The particular demographic variables shown (sex, race/ethnicity, age, region, parental education, and size and type of community) were selected because (1) they are major variables in NAEP reports and (2) they reflect different types of divisions of the population which might have different levels of sampling variability.

The table shows that the design effects are predominantly larger than 1, indicating that standard variance estimation formulas will be generally too small, sometimes markedly so. Further, the distributions of design effects appear different for certain subgroups of the population. The sampling variability of a subgroup, relative to the entire sample, depends on, among other things, how that subgroup is spread throughout the sample and what (weighted) proportion of the total sample is accounted for by the subgroup. For example, the white subgroup of the race/ethnicity variable is fairly evenly spread throughout the sample and accounts for more than 75 percent of the total sample by weight. Consequently, the distribution of design effects for this subgroup closely resembles that of the total population.

The subgroups determined by sex and parental education are also fairly evenly spread across the sample. In these cases, however, a given subgroup is a smaller proportion of the total population. Consequently, any effects of cluster selection (students within schools) on the variance estimates should be reduced, relative to the total population, because there are fewer observations per cluster but roughly the same number of clusters. The result is a tendency for the design effects for these subgroups to be somewhat lower than those for the total.

Table 4

Distribution of Design Effects
by Demographic Subgroup
for the Cognitive Mathematics Items Given in 1986

Grade 11*

Group	LoQ	Median	HiQ	Max	Mean
TOTAL	1.28	1.58	1.95	3.56	1.65
MALE	1.19	1.47	1.84	4.59	1.55
FEMALE	1.19	1.44	1.76	3.76	1.51
WHITE	1.17	1.49	1.85	3.82	1.55
BLACK	1.03	1.28	1.62	3.32	1.37
HISPANIC	1.10	1.35	1.81	4.05	1.53
OTHER	0.98	1.33	1.81	5.23	1.49
<AGE	1.06	1.28	1.59	3.27	1.33
=AGE	1.20	1.47	1.84	4.03	1.56
>AGE	1.21	1.46	1.81	4.36	1.52
NE	0.95	1.30	1.85	4.16	1.44
SE	0.96	1.36	2.15	7.26	1.68
CENTRAL	1.00	1.36	1.94	4.53	1.52
WEST	1.09	1.69	2.54	8.02	1.99
RURAL	0.97	1.52	2.40	8.12	1.82
LOW MET	1.01	1.42	1.93	4.59	1.55
HI MET	1.04	1.31	1.72	4.25	1.41
BIG CITY	0.97	1.22	1.62	4.45	1.34
FRINGE	0.88	1.22	1.69	3.64	1.35
MED CITY	1.16	1.60	2.08	4.21	1.69
SMALL PL	1.10	1.46	1.93	4.64	1.59
<H.S.	1.14	1.39	1.71	3.14	1.46
GRAD HS	1.08	1.37	1.66	4.51	1.42
POST HS	1.02	1.28	1.60	3.46	1.35
GRAD COL	1.20	1.48	1.83	3.46	1.56
UNKNOWN	1.01	1.24	1.57	4.11	1.31

* Distributions based on 372 multiple choice items

The distributions of design effects by region have noticeably more variability (as measured by the inter-quartile range) than that of the total sample. This is because the partitioning of the entire sample into regions occurs at the PSU level so that a PSU is either entirely included or entirely excluded in the estimation of statistics at the regional level. Since the level of aggregation used for variance estimation purposes is at the PSU level, the estimated variances of regional level statistics are based on fewer degrees of freedom than are those of national level statistics. Consequently, the sampling variability of the regional level variance estimates must be larger. We shall return to this point later.

Overall, although the distributions of design effects are somewhat different by subgroup, they are similar enough to warrant the selection of an overall composite value which would be adequate for most purposes. This overall design effect could then be used to adjust the conventional variance estimates for any subgroup to a value near to that which would be obtained by the jackknife. In choosing a composite design effect, some consideration must be made about the relative consequences of overestimating the variance as opposed to underestimating the variance. For example, adopting the position that an overestimate of the variance is as severe an error as an underestimate leads to using a composite which is near to the center of the distributions of the design effects. Possible composites of this type are the mean and median design effects. In the current data, the mean design effect is 1.5, which is close to, but greater than, the median design effect of 1.4.

Alternatively, one can adopt the position that it is a graver error to underestimate the variability of a statistic than to overestimate it. For example, Johnson and King (1987) examine estimation of variances using design effects (among other techniques) under assumption that the consequences of an underestimate are three times as severe as those of an overestimate of the same magnitude. Assuming that the distribution of design effects is roughly independent of the jackknife variance, so that the size of a design effect does not depend on the size of the variance, and adopting a loss function which is a weighted sum of absolute values of the deviations of predicted from actual with underestimates receiving three times the weight of overestimates, produces the upper quartile of the design effects as the composite value. The value of this composite design effect is 1.9.

Table 4 indicated what sort of design effects are to be expected for the simple case of weighted percent correct statistics and showed that an acceptable composite design effect would be in the range of 1.4 to 1.9, depending on how conservative one wishes to be. It is important to get an indication of how well such a composite would work in more complex situations, such as the case of regression coefficients. Table 5 addresses this question for our example regression.

Table 5 shows the design effects for the four regression coefficients from the weighted regression of Mpv1 on BW, NMATH and BW*NMATH. Also included in the table are the weighted estimates of the coefficients, their

conventional standard errors, and their jackknife standard errors. The conventional standard errors were obtained from a standard weighted regression in which the weights were scaled to add to the total sample size of 10105, thereby more closely representing the true amount of information available in the construction of the standard errors relative to the standard errors which would have been produced if the weights were left in their original metric (with a sum of 1,060,564). We see from Table 5 that this scaling of the weights does not go nearly far enough in terms of appropriately representing the true sampling variability of the regression estimates.

Table 5

Conventional versus Jackknife Standard Errors for the Weighted Regression of Mpv1 on Race and Highest Math Course Taken

Coefficient		Estimate	Standard Error		Design Effect
			Conventional	Jackknife	
Intercept	(α)	252.80	.64	.88	1.87
BW	(β)	7.75	.64	.83	1.66
NMATH	(γ)	14.54	.21	.35	2.65
BW*NMATH	($\beta\gamma$)	1.28	.21	.30	1.99

The design effects for the regression coefficients range from 1.7 to 2.7, indicating that the conventional estimates of the standard errors are too small by a factor of 1.3 to 1.6. It is interesting to note that these design effects appear somewhat larger than the design effects for the percent correct statistics indicating a somewhat larger effect of clustering for the regression coefficients, particularly in the case of the slope of Mpv1 on NMATH.

It is possible to approximate the correct sampling variance estimates for the regression statistics (and for the percent correct statistics) by selecting a compromise design effect, rescaling the weight of each individual so that the sum of the scaled weights is equal to the effective sample size, defined as the sample size divided by the design effect, and then conducting a traditional weighted analysis using these scaled weights. In our example, a reasonable compromise design effect is 2. Scaling the weights so that they add to the effective sample size of 5052=10105/2 and then running a weighted regression produces the same coefficient estimates as before, but standard errors of .90,.90,.30, and .30, which are much closer to the jackknife values.

While this approach will handle the estimation of sampling variability, it does not address another key issue: the degrees of freedom of the variance estimates.

1.3 The Degrees of Freedom of the Variance Estimate

It is important to have an indication of the number of degrees of freedom to attribute to the jackknife variance estimator $\hat{Var}(t)$. The degrees of freedom of a variance estimator provide information on the stability of that estimator: the higher the number of degrees of freedom, the lower the variability of the estimator. In practical terms, the number of degrees of freedom of the variance estimator corresponds to the number of residual degrees of freedom that can be assumed for inferential procedures.

Since the jackknife procedure estimates the sampling variability of the statistic by assessing the effect of change in the sample at the paired FSSU level, the number of degrees of freedom of the variance estimator $\hat{Var}(t)$ will be at most equal to the number of FSSU pairs. The number of degrees of freedom equals the number of independent pieces of information used to generate the variance. In the current case, the pieces of information are the 38 squared differences $(t_i - t)^2$, each supplying at most one degree of freedom (regardless of how many individuals were sampled within any FSSU).

The number of degrees of freedom of the sample variance estimator can be strictly less than the number of FSSU pairs. For example, suppose that the statistic t is a mean for some subgroup and no members of that subgroup can come from either FSSU in the i^{th} FSSU pair. (Examples of such a subgroup are any PSU-level partitioning of the population, such as region.) In this instance, neither member of the FSSU pair i directly contributes to the estimate of t, so that the component $(t_i - t)^2$ is measuring the effect of the post-stratification adjustments of the replicate weights. While being nonzero, such a component is likely to be smaller in magnitude than the squared difference $(t_k - t)^2$ for any PSU pair k which does contribute to the estimate of t.

In general, the squared difference $(t_i - t)^2$ will be estimating the variance component σ_i^2, say, which is the contribution to the sampling variance of the statistic t which can be attributed to the samples within the i^{th} FSSU pair. That is, $\hat{Var}(t)$ is estimating

$$\sum_{i=1}^{38} \sigma_i^2 .$$

If a few of the σ_i^2 are markedly larger than the remainder, as in the above case, then $\hat{Var}(t)$ is predominantly estimating the sum of these larger components which dominate the remaining terms. The effective degrees of freedom of $\hat{Var}(t)$ in this case will be nearer to the number of dominant terms.

It is possible to estimate how many degrees of freedom to attribute to the jackknife variance estimates of the weighted percent correct statistics

by comparing the distribution of design effects for a given population (such as males or total) with the quantiles of a chi-square distribution, choosing as the effective degrees of freedom the degrees of freedom of that chi-square distribution which most nearly matches the distribution of the design effects. More specifically, we assume that the jackknife variance estimate, V_j, of the j^{th} weighted percent correct statistic, P_j, is distributed like the random variable $(\sigma_j^2 / f)\, X_f^2$, where X_f^2 is a chi-square random variable with f degrees of freedom and σ_j^2 is the expected value of V_j. Then, for a sufficiently large sample size, and assuming that the expected value of the conventional variance estimate is σ_j^2/c, where c does not depend on j, the design effect of P_j will be approximately distributed like the constant (c/f) times a chi-square random variable with f degrees of freedom.

The value of the effective degrees of freedom for the design effects of a given population (such as males) is found by comparing the ordered values of the 372 design effects for that group with the expected values of the order statistics of a sample of 372 chi-square random variables with f degrees of freedom. The value of f such that the correlation between the ordered design effects and the chi-square order statistics is highest is chosen as the effective degrees of freedom for the set of design effects.

Table 6 shows the result of this estimation of the effective degrees of freedom of the design effects, and hence the jackknife variance estimates, for the weighted percent correct statistics. The quality of the approximation to the distribution of the design effects is quite good, in every case, the correlation between the ordered values of the design effects and their predicted values (the order statistics from a chi-square distribution with the effective degrees of freedom) exceeds .97.

The numbers in Table 6 show that the effective degrees of freedom of the jackknife variance estimates are indeed no larger than the number of FSSU pairs, and are, in fact, markedly smaller in all cases. There appear to be three major groupings:

1. Variance estimates for weighted percent statistics corresponding to subpopulations which are distributed fairly uniformly across the sample. These subpopulations are: total, the two sexes, Whites and Blacks, the 3 age groups, and the 5 levels of parental education. The variance estimates for these subpopulations have the highest number of effective degrees of freedom, all being around 20.

2. Variance estimates corresponding to subpopulations which are more concentrated in some portions of the sample than others, but which are represented to some extent throughout the sample. Examples of these are

Hispanics and Other Race/Ethnicity. These variance estimates have smaller degrees of freedom of around 10 to 13.

3. Variance estimates corresponding to subpopulations which can only appear in certain parts of the sample; such as the 4 regions and the 7 size and type of community designations. These have the smallest effective degrees of freedom, particularly so for the regions, which have between 5 to 9 effective degrees of freedom.

Table 6

Effective Degrees of Freedom for the Design Effects of the Percent Correct Statistics

Group	Effective D.F.
TOTAL	20
MALE	19
FEMALE	21
WHITE	18
BLACK	18
HISPANIC	13
OTHER	10
<AGE	23
-AGE	20
>AGE	21
NE	9
SE	6
CENTRAL	8
WEST	6
RURAL	5
LOW MET	10
HI MET	15
BIG CITY	15
FRINGE	10
MED CITY	10
SMALL PL	11
<H.S.	21
GRAD HS	19
POST HS	19
GRAD COL	21
UNKNOWN	17

It is also interesting to determine how many degrees of freedom to attribute to the variance estimates for the regression coefficients in our example regression. Table 7 provides this information. In constructing this table, a slightly different procedure was used from that used for Table 6 since, rather than having a distribution of design effects, we have only a single jackknife variance for each of the regression coefficients. In order to produce a distribution of variances for each of the coefficients, the full sample of 10105 respondents was split into 10 interpenetrating subsamples of 1011 respondents each by sorting the full sample by PSU and then systematically selecting the first student and every tenth student thereafter to be in the first subsample; the second student and every tenth student thereafter to be in the second subsample; and so on. The result is 10 samples, each of which include all of the PSUs in the full sample while including 1/10 of the students within each of the PSUs. As we noted above, the jackknife variance estimate based on one of these subsamples will be larger, because of the reduced sample size, but will have the same number of degrees of freedom as did the estimator from the full sample since the same number of FSSU pairs are involved.

Jackknife variance estimates for the regression coefficients were estimated for each of the 10 subsamples and then, for each regression coefficient, the best fitting chi-square distribution to the distribution of the variance estimates was found, in a manner analogous to that used for the design effects. The resulting effective degrees of freedom for the four regression coefficients appear in Table 7. As was the case for the design effects, the approximation is quite good, with the correlation between the ordered values of the variance estimates and their predicted values (the order statistics from a chi-square distribution with the effective degrees of freedom) exceeding .96 for each regression coefficient.

Table 7

Effective Degrees of Freedom for the Jackknife Standard Errors of the Weighted Regression of Mpvl on Race and Highest Math Course Taken

Coefficient		Effective D.F.
Intercept	(α)	28
BW	(β)	22
NMATH	(γ)	22
BW*NMATH	($\beta\gamma$)	15

We see again that the effective degrees of freedom of the variance estimates is smaller than the number of FSSU pairs, and that this is true for each of the coefficients. Furthermore, the number of degrees of freedom varies with the type of coefficient with the largest number, 28, attributed to the intercept, a smaller value, 22, to the BW effect and the slope on

NMATH, and the smallest value, 15, attributed to the BW * NMATH interaction. The relation between the effective number of degrees of freedom and the type of regression coefficient is related to the sensitivity of the coefficient's variance estimate to the clustering across the sample of the observations going into the estimate. This sensitivity is determined by the most clustered subgroup which determines the regression coefficient. Thus, the sensitivity of the BW effect is determined by the clustering of the Black students, the sensitivity of the NMATH slope by the clustering of students within any one of the levels of NMATH, and the sensitivity of the interaction by the clustering in a two-way cell of Black students at a given level of NMATH.

The effective degrees of freedom for the NAEP jackknife variance estimates are markedly smaller than the degrees of freedom attributed to the corresponding error estimates from conventional techniques. This fact affects inferential procedures since significance tests based on the conventional degrees of freedom will be too liberal (and confidence intervals will be too small). Fortunately, for the usual significance levels, the effect of using the effective degrees of freedom rather than the conventional values is generally moderate: a t statistic significant at the α=5% level assuming infinite degrees of freedom (essentially the conventional estimate) is significant at the α=6% level for 20 effective degrees of freedom, the α=7% level for 10 effective degrees of freedom, and the α=10% level for 5 effective degrees of freedom.

For practical purposes, the impact of the reduced degrees of freedom on inferential techniques can be largely accounted for by (1) using a moderate number (say 25) of degrees of freedom for all inferences, and (2) being cautious about barely significant results for subgroups which are highly clustered in the population.

2. THE EFFECT OF MEASUREMENT ERROR ON ANALYSIS

The previous discussion has considered the effects of the sample design on analysis when the variable of interest is observed without error on every respondent. This is the case for statistics such as weighted percents-correct, but is not the case for the scale-score proficiency values, such as the Mathematics composite, used by NAEP. The NAEP scale-score proficiency values are not test scores in the usual sense. Traditional IRT applications provide point estimates of individual proficiencies; such point estimates are not warranted for the NAEP data because of the relatively small number of items per subarea answered by any individual.

The basis for the measurement of mathematics skills is a pool of 372 cognitive items. Although these items formed the basis of the mathematics proficiency subscales which were created in 1986, it is not necessary or even desirable that each assessed student respond to the entire battery of items. Firstly, such a request would require a substantial amount of each student's time (nearly 3 hours), to the extent that the response rates on the part of students and of schools would be expected to be deleteriously affected. Secondly, the goal of an assessment is to estimate aggregate performance of provide measures of subgroups of the population rather than to obtain precise measures for individuals. This means that it is not necessary that each student take all items and, in fact, for the goals of the assessment, it is more efficient in a statistical sense if each person takes only a selected subset of the items.

The IRT models used by NAEP to summarize performance in a learning area, such as Mathematics, assume the existence of a vector $\underline{\theta}$ of unobservable proficiencies. Each of the elements of the vector $\underline{\theta}$ corresponds to one of the subscales defined for Mathematics and summarizes the performance over the items within the subarea which determines the subscale. To provide an overall measure of Mathematics achievement, NAEP has also defined a Mathematics composite, which is a weighted average of the subscale proficiencies. Because each individual typically takes few items within any subarea, the values of $\underline{\theta}$, and the value of the composite, are not well determined for any individual.

That is, there exists a nontrivial amount of error of measurement in the estimated proficiency values for any given student. In order to account for this error of measurement, we have adapted Rubin's (1987) multiple imputation procedures for missing data to the estimation of latent variables. The essential idea is that, while we do not know precisely the proficiency of any individual, we can estimate the distribution of <u>potential</u> proficiency values for that individual. This estimated distribution takes into account the responses of the student to all Mathematics items presented to that student as well as answers to certain background and demographic questions.

To correctly estimate the value of any function $T(\theta)$ of some proficiency θ, such as, for example, the mean of the Mathematics composite, it is necessary to account for the uncertainties imposed by imperfectly estimating the proficiency value of each student. This is accomplished by taking the distribution of the potential values of proficiency for each student into account by, in effect, evaluating the expectation of $T(\theta)$ over the joint distribution of potential proficiencies for all assessed students. The exact evaluation of this expectation requires numerical integration. To avoid the need for this, a set of "plausible values" is created for each student by drawing values at random from the distribution of potential proficiency scale-scores for that student.

These plausible values are used to estimate the statistic $T(\theta)$ and its variance in the following manner. Let $\underline{\theta}_i^*$ be the vector consisting of the i_{th} plausible value for each of the sampled students. This vector is a

plausible representation of what the vector $\underline{\ell}$ of true proficiencies might have been, had we been able to observe them. Let t_i be the value of T(.) evaluated using $\underline{\ell}_i^*$ as if those plausible values were the true proficiencies. The statistic t_i provides a good estimate of the function $T(\underline{\ell})$, but does not individually indicate the effect on the estimate of not really knowing $\underline{\ell}$. An even better estimate of $T(\underline{\ell})$ is the average t^* of M (>1) such estimates t_i, where each of the estimates is based on a different set of the plausible values. (NAEP typically uses M=5 for the number of replicates of the plausible values).

The variance of t^* has two components, one due to the fact that a sample of students was used in its estimation, and the other due to the error of measurement of the proficiencies of the sampled students. The component of variance due to sampling variability can be estimated by applying the jackknife variance estimator in turn to each of the statistics t_i, denoting the jackknife variance of t_i by U_i. The overall best estimate of the sampling variance of t^* is then the average, U^*, of the M sampling variances of the t_i. (Actually, for computational expediency, NAEP uses a single jackknife estimate, U_1, in place of the average of M).

To measure how well t^* approximates $T(\underline{\ell})$ given the error of measurement, we assess the dependence of the value of the statistic on the particular set of plausible values used in its computation. This "error due to imputation" is estimated by the variance among the M estimates:

$$B_M = \sum_{i=1}^{M} (t_i - t^*)^2 / (M-1).$$

The final estimate of the variance of t^* is the weighted sum of these two components:

$$V = U^* + (1 + M^{-1}) B_M .$$

(The factor $(1 + M^{-1})$ before B_M accounts for the fact that only a finite number of plausible values were used in estimating the error due to imputation).

For the purposes of inference, the distribution of $(t^* - T)/V^{1/2}$ is assumed to be approximately t, with degrees of freedom given by

$$\nu = \kappa (M - 1) (1 + r_M^{-1})^2 ,$$

where r_M is the relative increase in variance due to not observing the $\underline{\ell}$ values:

$$r_M = (1 + M^{-1}) B_M / U^* ,$$

and κ is a function of r_M and d, the effective degrees of freedom of U^* :

$$\kappa = d / (d + (M-1)\, r_M^{-2}) .$$

The value of ν always exceeds the degrees of freedom, M-1, of B_M and is close to the degrees of freedom, d, of U^* whenever B_M is small relative to U^* so that the effect of the error of measurement is small relative to the sampling variability. (In fact, if B_M is sufficiently small, the value of ν can exceed d).

Rubin provides an index to describe the effect of not observing the missing data in the form of a measure of the fraction of information about T that is missing because $\underline{\theta}$ was not observed, relative to the information about T that would have been available if $\underline{\theta}$ were observed:

$$\gamma_M = \frac{r_M + 2/(\nu+3)}{r_M + 1} .$$

The value of $1-\gamma_M$ is roughly analogous to reliability in classical test theory and represents the degree of information about an average respondent's θ conveyed by that student's observed responses.

As an example of the effect of the measurement error on estimation, we return to our example regression, repeating the estimation of the regression coefficients, in turn, for each of the 5 plausible values provided for each student. Table 8 shows the result. We see that there is some variation in the estimates of the coefficients when these are based on different plausible values, particularly so for the BW effect.

Table 8

Coefficients of the Weighted Regression of Mathematics Proficiency on Race and Highest Math Course Taken Based on each of the 5 Plausible Values

Coefficient		Plausible Value				
		# 1	# 2	# 3	# 4	# 5
Intercept	(α)	252.80	253.28	253.72	253.76	253.87
BW	(β)	7.75	7.53	6.59	7.01	6.09
NMATH	(γ)	14.54	14.30	14.26	14.25	14.24
BW*NMATH	($\beta\gamma$)	1.28	1.38	1.66	1.50	1.74

Table 9 shows the values of the statistics U^*, d, B_M, V, r_M, ν, and γ for each of the regression coefficients.

Table 9

Estimated Error Variances and Related Coefficients for the Regression Coefficients

Coefficient		U^*	d	B_M	V	r_M	ν	γ_M
Intercept	(α)	.772	28	.195	1.006	.303	29	.28
BW	(β)	.687	22	.460	1.239	.804	16	.51
NMATH	(γ)	.121	22	.017	.141	.163	26	.20
BW*NMATH	($\beta\gamma$)	.091	15	.036	.134	.472	18	.39

We see that r_M, the relative increase in the variance of the regression coefficients due to the error of measurement of the proficiencies, is fairly substantial, particularly in the case of the BW effect where the overall variance V=1.239 is 80% larger than the its sampling variance U^*. This increase is due to the substantial amount (more than half) of the information about that coefficient which is missing due to the latency of the proficiencies. The other coefficients also exhibit marked increases in the total variances, relative to their sampling variances, but not to the same degree as BW, since relatively less information is missing for these other regression coefficients.

The plausible values technology has provided an honest indication of how well the regression coefficients have been measured and tells us, particularly in the case of the BW effect, that there is considerable variability in their measurement.

SUMMARY

This paper has considered the effects of certain characteristics of the NAEP data on analysis. It was found that:

1) the bias due to ignoring the sampling weights was often moderate but that weighting becomes more important in more complex situations,

2) the effect of the sample design on variance estimation is to roughly double the size of the variance estimates relative to the conventional estimate,

3) the effective degrees of freedom of the jackknife variance estimates are much smaller than the conventional estimates, and

4) the effect of the measurement error in the estimation of functions of scale-score proficiency values can be substantial and must be accounted for in analysis.

Overall, the consequences of ignoring the special features of the NAEP data and proceeding with a standard analysis, is to produce inferences which are too liberal since they underestimate the true variability and overestimate the true degrees of freedom. Fortunately, it is possible to account for the effects of the design, and the latency of the proficiency values, by using the jackknife and plausible values technologies. The price to be paid in using these technologies is an increase in the amount of computation required; a price which is not too high considering the increased validity of the final result.

REFERENCES

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1953). Sample Survey Methods and Theory, Volume 1. New York: John Wiley & Sons.

Johnson, E. G. & King, B. F. (1987). Generalized Variance Functions for a Complex Sample Survey. Journal of Official Statistics, 3, 235-250.

Kish, L. (1965). Survey sampling. New York: John Wiley & Sons.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. Journal of the Royal Statistical Society, Series B, 36, 1-22.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons.

Wolter, K. M. (1985). Introduction to Variance Estimation. New York: John Wiley & Sons.